bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Decomposing spatially dependent and cell type specific contributions to cellular heterogeneity

Qian Zhu[1], Sheel Shah[2,3], Ruben Dries[1], Long Cai[2]*, Guo-Cheng Yuan[1]*

1. Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T. H. Chan School of Public Health, Boston, MA 02215, USA
2. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
3. UCLA-Caltech Medical Scientist Training Program, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095, USA
*Co-corresponding authors: lcai@caltech.edu (L.C.); gcyuan@jimmy.harvard.edu (G.C.Y.)

Abstract

Both the intrinsic regulatory network and spatial environment are contributors of cellular identity and result in cell state variations. However, their individual contributions remain poorly understood. Here we present a systematic approach to integrate both sequencing- and imaging-based single-cell transcriptomic profiles, thereby combining whole-transcriptomic and spatial information from these assays. We applied this approach to dissect the cell-type and spatial domain associated heterogeneity within the mouse visual cortex region. Our analysis identified distinct spatially associated signatures within glutamatergic and astrocyte cell compartments, indicating strong interactions between cells and their surrounding environment. Using these signatures as a guide to analyze single cell RNAseq data, we identified previously unknown, but spatially associated subpopulations. As such, our integrated approach provides a powerful tool for dissecting the roles of intrinsic regulatory networks and spatial environment in the maintenance of cellular states.

Introduction

Human and other multicellular organisms are composed of diverse cell types characterized by distinct gene expression patterns. Within each cell type, there is also considerable heterogeneity. The source of cellular heterogeneity remains poorly understood, but it is commonly thought to be modulated by the balance between intrinsic regulatory networks and extrinsic cellular microenvironment (Swain et al., 2002; Jaenisch and Bird 2003). Recently, the rapid development of single-cell technologies has enabled accurate and simultaneous measurements of cell position and gene expression (Yuan et al. 2017), thus providing an

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

excellent opportunity to systematically dissect the differential roles of intrinsic and extrinsic factors on mediating cellular heterogeneity.

Currently, there are two major, complementary approaches for single-cell transcriptomic profiling. The first is single-cell RNA sequencing (scRNAseq) (Tang et al. 2009; Islam et al. 2011; Dalerba et al., 2011; Deng et al., 2014; Jaitin et al., 2014; Macosko et al. 2015; Klein et al 2015). By combining single-cell isolation, library amplification, and massively parallel sequencing, scRNAseq provides the most comprehensive view of transcriptomes. The second approach is single-molecule fluorescence in situ hybridization (smFISH) (Raj et al., 2008; Lubeck and Cai, 2014; Chen et al., 2015; Moffitt et al. 2016; Shah et al. 2016a; Shah et al. 2016b), which can be used to detect mRNA transcripts with high sensitivity while maintaining the spatial content. With sequentially rounds of smFISH imaging, it is now feasible to profile the expression level of hundreds of genes for each cell in tissues. Each technology features a distinct set of advantages and limitations. The sequential FISH technology carries the advantage of measuring the transcriptome with high accuracy in its native spatial environment, but current implementations profile only a few hundred genes, whereas single-cell RNAseq provides whole-transcriptome estimation but requires cells to be removed from their environment, resulting in a loss of spatial information.

It is clear that an integrative analysis framework, involving single-cell RNAseq and sequential FISH, would bring together the benefits of both technologies to better characterize both cell type and spatially dependent variations. To this end, we developed a computational approach that contains two major components: First, the single cell RNAseq data is used as a guide to accurately determine the cell-types corresponding to the cells profiled by sequential FISH. Second, distinct spatial domain patterns are systematically detected from sequential FISH data. These spatial patterns are then in turn used to dissect the environment-associated variation in a single-cell RNAseq dataset.

This integrated approach has enabled us to systematically dissect the respective contribution of cell type and spatially dependent factors in mediating cell-state variation (**Fig. 1a**), which has eluded previous studies. Most existing studies focused on identifying cell-type differences, but, as shown below in our analysis of the mouse visual cortex region, cell-type differences represent only one component in cell-state variation (schematically represented as the cell intrinsic dimension in **Fig. 1a**), whereas local environment plays a significant role in mediating gene activities, probably through cell-cell interactions (represented as the spatial dimension in **Fig. 1a**) and signaling. As each technology has its own strengths and weaknesses, the integrated approach presented here provides a powerful model framework and broadly applicable to analyze diverse tissues from various model systems.

Results

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Mapping scRNAseq cell-types on seqFISH data

Given that scRNAseq, as a whole transcriptomic approach, can provide signatures for a diverse set of cell types, we took advantage of the whole-transcriptomic information obtained from scRNAseq data (Tasic et al., 2016) and developed a supervised cell-type mapping approach by integrating seqFISH and scRNAseq data (**Fig. 1b**). Our goal differs from previous studies (Achim et al., 2015; Satija et al., 2015; Halpern et al., 2017; Karaiskos et al., 2017), where scRNAseq data were mapped onto conventional ISH images to predict cell locations. Of note, ISH images are not quantitative, multiplexed or single-cell resolution. In a seqFISH experiment, transcripts from hundreds of genes are detected directly in individual cells in their native spatial environment at single molecule resolution.

Our strategy is to use scRNAseq data to capture the large cell type differences and then further investigate spatial patterning within each major cell type. We analyzed a published scRNAseq dataset targeting the mouse visual cortex regions (Tasic et al., 2016). Eight major cell types: GABAergic, glutamatergic, astrocytes, 3 oligodendrocyte groups, microglia, and endothelial cells were identified from scRNAseq analysis (Tasic et al., 2016). To estimate the minimal number of genes that is required for accurate cpe mapping, we randomly selected a subset from the list of differentially expressed (DE) genes across these cell types, and applied a multiclass support vector machine (SVM) (Cortes and Vapnik, 1995; Fan et al., 2008) model using only the expression levels of these genes. The performance was evaluated by cross-validation. By using only 40 genes, we can already achieve an average level of 89% mapping accuracy. Not surprisingly, increasing the number of genes leads to better performance (92% for 60 genes, and 96% for 80 genes). Therefore, there is significant redundancy in transcriptomic profiles which can be compressed into fewer than 100 genes.

We then investigated a seqFISH dataset for the mouse visual cortex area (Shah et al., 2016a). A 1 mm by 1 mm contiguous area of the mouse visual cortex was imaged with 4 barcoded rounds of hybridization to decode 100 unique transcripts followed by 5 rounds of non-combinatorial hybridization to quantify 25 highly expressed genes (**Supplementary Table 1**). These rounds of imaging were preceded by imaging of the DAPI stain in the region and followed by imaging of the Nissl stain in the region. The images were aligned and transcripts decoded as described in Shah et al. 2016. Transcripts were assigned to cells that were segmented based on Nissl and DAPI staining. Using this technology, we were able to quantify the expression levels of these 125 genes with high accuracy in a total of 1597 cells.

After computing differentially expressed genes across the 8 major cell types in (Tasic et al, 2016), we selected the top 43 (P<1e-20) of these 125 genes for cell-type classification. These genes contain both highly expressed (>50 copies per cell) and lowly expressed genes (<10 copies per cell). Cross-validation analysis shows that, using these 43 genes as input, the SVM

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

model accurately mapped 90.1% of the cells in the scRNAseq data to the correct cell-type. Therefore, we proceeded by using these 43 genes (**Supplementary Table 2**) to map cell-types in the seqFISH data.

As a first step, we preprocessed the seqFISH data by using a multi-image regression algorithm in order to reduce potential technical biases due to non-uniform imaging intensity variation (Methods). We further adopted a quantile normalization (Bolstad et al., 2003) approach to calibrate the scaling and distribution differences between scRNAseq and seqFISH experiments. For most genes, the quantile-quantile (q-q) plot normalization curve is strikingly linear (**Supplementary Fig. 1**), suggesting a high degree of agreement between the two datasets despite technological differences.

Then, the SVM classification model was applied to the bias-corrected, quantile-normalized seqFISH data to assign cell types. Of note, we found that better performance may be achieved by further calibrating model parameters to accommodate platform differences. The results of multiclass SVM are calibrated across models (Platt, 1999) and converted to probabilities. The results showed the exclusion of 5.5% cells that cannot be confidently mapped to a single cell-type (with 0.5 or less probability). Among the mapped cells, 51% are glutamatergic neurons, 35% are GABAergic neurons, 4.5% are astrocytes, and other glial cell types and endothelial cells make up the remaining 4% of cells (**Fig. 1c**).

To validate our predictions, we first checked the expression of known marker genes and compared the average gene expression profiles between scRNAseq and seqFISH data. Indeed, this comparison shows a high degree of similarity (**Fig. 1c**). Notably, marker genes have expected high expression in the matched cell types, such as Gja1 and Mfge8 in astrocytes, Laptm5 and Abca9 in microglia, Cldn5 in endothelial cells, Tbr1 and Gda in glutamatergic neurons, and Slc5a7 and Sox2 in GABA-ergic neurons. The majority of cell types have a high Pearson correlation (>0.8) between matched cell types' average expression profile; even for the rare cell-type microglia, the correlation remains reasonably high (0.75) (**Fig. 1d**). We are also able to distinguish early maturing oligodendrocytes in the seqFISH data based on Itpr2 expression (**Fig. 1c,** OPC.1 column) as previously reported (Zeisel et al, 2015). Inhibitory GABA-ergic neurons and excitatory glutamatergic neurons exhibit strong anti-correlation to each other (**Fig. 1d**).

As an additional validation, we examined the Nissl and DAPI staining images which are known to have distinct patterns between astrocytes and neuronal cell types. As Nissl is a neuronal stain and DAPI stains DNA, astrocytes are typically associated with DAPI but not Nissl, whereas neurons are stained for both. Our cell-type mapping results highly agree with these patterns. Over 89% of predicted astrocytes exhibit strong DAPI staining but weak or no Nissl staining

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

across cortex columns (**Supplementary Fig. 2**, **Supplementary Table 3**). Taken together, these analyses strongly indicate that the vast majority of cells were mapped to the correct cell types.

By combining cell type predictions from scRNAseq and positional information from seqFISH, we were able to construct a single-cell resolution landscape of cell type spatial distribution (**Fig. 1e**). As expected, this landscape is very complex, with different cell types intermixed with each other (**Fig. 1e**). On the other hand, it is clear that there remains significant heterogeneity within each cell-type.

A systematic approach to identify multicellular niche from spatial genomics data

Microenvironment in tissues can contribute to heterogeneity in addition to cell type specific expression patterns. To systematically dissect the contributions of microenvironments on gene expression variation, we developed a novel hidden-Markov random field (HMRF) approach (Zhang et al., 2001) to unbiasedly inform the organizational structure of the visual cortex. An overview of this approach is illustrated in **Fig. 2a**. The basic assumption is that the visual cortex can be divided into domains with coherent gene expression patterns. A domain may be formed by a cluster of cells from the same cell-type, but it may also consist of multiple cell-types. In the latter scenario, the expression patterns of cell-type specific genes may not be spatially coherent, but environmentally associated genes would express in spatial domains. HMRF enables the detection of spatial domains by systematically comparing the gene signature of each cell with its surroundings to search for coherent patterns. Briefly, we computationally constructed an undirected graph to represent the spatial relationship among the cells, connecting any pair of cells that are immediate neighbors (**Fig 2a, b**). Each cell is represented as a node in this graph. The domain state of each cell is influenced by two sources (**Fig 2b**): 1) its gene expression pattern, and 2) the domain states of neighboring cells. The total contribution of neighboring cells can be mathematically represented as a continuous energy field, and the optimal solution is identified by searching for the equilibrium of the energy field (see Methods for mathematical details).

Next, we applied our HMRF model to analyze the 1597-cell mouse visual cortex seqFISH dataset. For the visual cortex region, the detection of spatial patterns is confounded by the fact that different cell types tend to be mixed together. To reduce this confounding effect, we systematically removed genes that are strongly associated with specific cell-types. We further narrowed down the gene list by identifying genes with spatially coherent gene expression patterns using a Silhouette metric (see Methods). This resulted a list of 69 genes (**Supplementary Table 4**) that were used to identify spatial domains.

HMRF modeling of the visual cortex region revealed 9 spatial domains (**Fig. 2c**). These domains have distinct spatial patterns; some display a layered organization that resembles the

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

anatomical structure (Sunkin et al., 2013). For example, four of the domains are located on the outer layers of the cortex therefore labeled as O1, O2, O3, and O4, respectively (**Fig. 2c**). The locations of these layers roughly correspond to the well-characterized L1, L6, and external capsule (EC) layers, respectively. Four domains are located on the inside of the cortex therefore labeled as I1a, I1b, I2, and I3, respectively (**Fig. 2c**). These domains roughly correspond to the L2-5 layers. These inner domains are less pronounced than the outer domains, which is consistent with previous anatomical analysis. Finally, one domain is sporadically distributed across in the inner layers of the cortex, therefore labeled as IS (**Fig 2c**).

By overlaying cell type annotations, we see that each domain generally consists of a mixture of GABA-ergic, glutamatergic neurons and astrocytes interacting in each environment (e.g. domain I1a in **Supplementary Fig. 3**). We further revisited Nissl staining to observe the physical characteristics of these cells in HMRF-defined domain environment. Strikingly, the domains identified by HMRF correspond very well with distinct shapes of the cells in the outer layer domains O2, O3, O4, which exhibit the characteristics of elongated cells, small size, large circular cells respectively (**Fig 2c**). Some of these differences are cell-type related. However, very often within a cell-type, such as glutamatergic neurons, there remains significant morphological differences across domains, as described in the next section, suggesting that spatial positions accounts for a large part of morphologies in these cells, consistent with known morphological diversity in the cortex. Overall, cells located in different HMRF domains are associated with distinct morphologies.

The decomposition of mouse visual cortex into spatial domains suggests that a spatial gene expression program is shared across cells in proximity. Differential gene expression analysis identified distinct signatures associated with each spatial domain (**Fig. 3a**). For example, genes Calb1, Cpne5, Nov are preferentially expressed in inner domains (I1a, I1b), whereas genes Tbr1, Serpinb11, Capn13 are highly enriched in outer domains (O1, O2). Different outer domains can be further distinguished by additional markers, such as Mmp8 (O2), Spag6 (O1), and Neurod4 (O1). The spatial marker genes are highly consistent with their spatial expression in Allen Brain Atlas (Sunkin et al., 2013) ISH images, such as Calb1, Cpne5, Nov, Gda, and Tbr1 (see **Supplementary Figs. 4, 5**). Additional genes such as Nell1, Aldh3b2, Gdf5 are also consistent with cell clusters in an independent dataset (Zeisel et al. 2015) (**Supplementary Fig. 5**). Taken together, these analyses strongly suggest that our model for analyzing seqFISH data is able to detect functionally, morphologically, and transcriptionally distinct spatial environments.

Integrative analysis identified cell-type, environmental interactions

Glutamatergic neurons mediate the neuronal circuit in the visual cortex by playing a primarily excitatory function. It is also well-known that the behavior of different glutamatergic neurons can be very different (Andjelic et al., 2008; Tasic et al., 2016). By combining cell-type mapping

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

and spatial domain identification, we set out to dissect the source of heterogeneity within glutamatergic cells.

First, nearly all glutamatergic cells express cell-type specific markers such as Tbr1 and Gda (**Fig 3b top**). In addition to demonstrating cell type identity, there exists substantial heterogeneity within glutamatergic cells in a spatially dependent manner. As glutamatergic cells are spread across all 9 domains, each subset expresses a different gene signature in accordance to domain annotation (**Fig. 3b middle**). Furthermore, an additional set of gene signatures are differentially expressed between glutamatergic cells in different domains (**Fig. 3b bottom**). For example, Neurog1 in domain IS, is a IS-domain specific gene upregulated in glutamatergic cells but not in GABA-ergic neurons or other cell types (**Fig 3b bottom**). Other genes such as Vmn1r65, Psmd5, follow a similar specific pattern (**Fig 3b bottom**). Collectively, the domain-specific signatures map out the spatial patterns of expression within glutamatergic cells, demonstrating their power to differentiate subgroups of this cell type (**Supplementary Fig. 6**). Additionally, these spatially dependent variations within glutamatergic neurons have strong support from cell morphology. We compared the morphology of cells at the boundary of two layers for six different snapshot regions (**Fig 3c**). In every case, domain boundaries clearly mark the boundaries of layers that possess visually identifiable cell shape characteristics (the three groups of cells in panel L6a, L6b, EC of **Fig 3c**). Therefore, glutamatergic cells in different domains show striking morphological differences, further supporting the validity of our domain partition results. Together, these analyses strongly suggest that spatial domain variation plays an important role in mediating cellular heterogeneity within a common cell-type.

Using HMRF domain information to reanalyze scRNAseq data

Single-cell RNAseq data does not contain spatial information. However, by integrating information from seqFISH data analysis, we were able to identify metagene signatures associated with different spatial domains. Briefly, for each domain we defined a metagene signature representing the gene set that is specifically expressed in this domain (see Methods). Using these metagenes as a guide, we were able to infer the spatial location of a cell based on the activities of these metagenes, defined as the average gene expression level. We used this approach to dissect the contribution of environmental factors to transcriptomic heterogeneity within glutamatergic cells from single-cell RNAseq data. t-SNE and k-means clustering analyses revealed a landscape of subpopulations associated with distinct metagene activities, which was strikingly consistent with seqFISH data analysis (**Fig 4a, b**).

For simplicity, these clusters were labeled according to their enriched metagene signatures. We identified differentially expressed genes in the single-cell RNAseq data between the aforementioned clusters, and examined their biological functions by using gene set enrichment

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

analysis. Interestingly, different biologically processes were associated with different domains, suggesting functional differences between the spatial domains (**Fig 4d**).

Importantly, subpopulations detected by metagene analysis are well enriched in the manual layer annotations provided from authors of the dataset (**Fig 4c**). For example, cluster 5 (annotated as domain I1b based on metagene analysis) is enriched in L1-L2/3 dissected cells from Tasic et al ($P<1.2e-6$) (**Fig 4c**). Cluster 4 (marked as domain O1) is enriched in L6b dissection labels ($P<0.0017$). Cluster 9 (marked as domain IS) is enriched in L4 dissection label ($P<0.016$). Overall, these results demonstrate the value of our analysis in reinterpreting the scRNAseq dataset by mapping our spatial HMRF-derived signatures to RNAseq which contains no spatial information. Thus, integrating seqFISH data analysis provides new insights into scRNAseq data.

HMRF analysis reveals region-specific variation among astrocytes

Next, we investigated the environment effect on astrocytes, which are also known to have substantial heterogeneity (Ben Haim and Rowitch, 2016). Our cell type mapping identified 47 astrocytes in the seqFISH data. These cells all expressed key astrocyte markers (**Fig 5a, box 1**) but were spread across 5 HMRF domains (O1, O2, O3, I1a, and I3) (**Fig. 5a**). Furthermore, it is notable that several groups of environment associated genes are identified, indicative of key environmental processes. These signature genes are confirmed to be expressed in astrocytes according to bulk astrocyte RNAseq database (Zhang et al., 2016) (**Supplementary Fig. 7**). As an example, Sox2 and loxl1 in our domain I1a are two of most highly ranked astrocyte genes in bulk sequencing. Coexpression of these genes with other ECM (extracellular matrix) markers in the same state, such as Acta2, Col5a1, implicate an important role of ECM in this domain of astrocytes, which has been previously linked to the differentiation and reprogramming of astroglial lineage (Niu et al., 2015). While these ECM genes are upregulated in domain I1a, in other domains such as outer O1, O2, they are notably absent or down-regulated. Therefore, domain-specific astrocytes gene expression may reveal functional differences in different microenvironments.

Conclusion

A major goal in single-cell analysis is to systematically dissect the contributions of cell-types and environment on mediating cell-state variability (Regev et al., 2017). To achieve this goal, we presented an HMRF-based computational approach to combine the strengths of sequencing and imaging-based single-cell transcriptomic profiling strategies. We showed that our method can be used to correctly detect spatial domains in the mouse visual cortex region. In doing so, we were able to identify environment-associated variations within a common cell-type. Our analysis also demonstrated that novel insights can be gleaned from single-cell data by an

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

integration of information from complementary technologies. In particular, integrating single-cell RNAseq data allows us to map cell-types more accurately than in seqFISH data analysis, whereas integrating seqFISH data allows us to extract spatial structure in single cell RNAseq data analysis. Future work will continue to investigate the mechanisms underlying the interactions between cell-type and microenvironment.

Author Contributions

Conception and supervision of project: G.C.Y., L.C. Conception of HMRF and SVM models: Q.Z., G.C.Y. Conducting and supervision of computational analyses: Q.Z, G.C.Y. Conducting and supervision of seqFISH experiments: S.S, L.C. Writing: Q.Z., S.S., R.D., G.C.Y., L.C. All authors contributed ideas for this work. All authors reviewed and approved the manuscript. This research was supported by a Claudia Barr Award and NIH grant R01HL119099 to G.C.Y. and NIH R01 HD075605 to L.C.

Methods

SeqFISH data generation

SeqFISH data in the mouse visual cortex region was generated as described previously (Shah et al., 2016a). Briefly, 100 genes were encoded using a temporal barcoding method and 25 genes were quantified individually. To encode 100 genes, 4 rounds of hybridization were performed using 5 distinct fluorescence channels. Out of a total possible 625 barcodes, 100 were chosen such that loss of signal in any given hybridization still allows accurate decoding of the spot. Every transcript was hybridized in every round using a given probe set. After hybridization, the signal was amplified using smHCR and images were taken at predefined locations in the mouse visual cortex. The DNA probes along with the amplification polymers were digested using DNase I leaving behind a naked RNA for re-hybridization with the next probe set. A round of imaging with DAPI staining was done before any RNA hybridization to image all nuclei in the fields and a final round of Nissl staining was imaged to identify cell boundaries. Cells were segmented based on DAPI staining, Nissl staining, and RNA point density. Once all imaging rounds were completed, these images were aligned using a 2D normalized cross correlation and each spot was decoded based on the unique color switching pattern. For the 25 genes that were labelled without any encoding, simple spot counting was done to identify the number of transcripts. These transcripts were then assigned to cells based on the location of the transcript and the segmentation masks. For a more details regarding the seqFISH method, please refer to Shah et al. 2016. The spatial coordinates of the cells are provided in **Supplementary Data**.

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

SeqFISH data normalization and bias correction

The seqFISH gene expression matrix, represented by -log (count + 1), was normalized by row and column z-scoring to remove cell-specific and gene-specific biases. Potential field imaging biases were estimated and removed by using a multi-image regression algorithm similar as previously done (Caicedo et al., 2017). Briefly, for each gene, the imaging bias at each binned location was estimated by averaging the normalized gene expression levels over 8 neighboring bins within each field followed by averaging across all fields. The estimated bias was then modeled by principal component analysis (PCA). The contributions of the four most significant PCs were estimated by linear regression and removed from the normalized gene expression matrix (**Supplementary Fig 8**).

Cell type mapping

Single-cell RNAseq data for the mouse visual cortex were obtained from Gene Expression Omnibus (GSE71585). Cell-type information corresponding to 1723 cells was obtained from the original paper (Tasic et al, 2016). In this analysis, we considered the 8 major cell types: GABAergic, glutamatergic, astrocytes, 3 oligodendrocyte groups, microglia, and endothelial cells. Differentially expressed genes among different cell types were identified by MAST (Finak et al., 2015).

We trained classifiers of cell types from single-cell RNAseq dataset by using the multiclass SVM formulation. For each cell-type, we built a classifier as follows. Let x_i, $i = 1, \ldots, n$, be the gene expression pattern for the i -th cell, and y_i code for cell-type identity: $y_i = 1$ if cell i belongs to the specified cell type and -1 otherwise. We selected the linear kernel that produces two hyperplanes that best separates the two classes. The objective function is defined as follows

$$minimize\ C\left(\sum_{i=1}^{n} \zeta_i^2\right) + \|w\|^2/2$$

$$subject\ to\ 1 - \zeta_i \leq y_i\left(w \cdot x_i - b\right), \zeta_i \geq 0 \qquad \text{Eq. 1}$$

Here *w* is the normal vector to the hyperplane used to represent margin. The squared hinge loss function $\sum_{i=1}^{n} {\zeta_i}^2$ is used here to quantify the margin of misclassification error. *C* is a regularization parameter that trades off misclassification due to overfitting against simplicity of the decision function. A lower *C* increases the ability of the model to generalize to unseen data at a cost of larger fitting error. For testing data, the sign of $w \cdot x_i - b$ is used to predict cell type identity. We used the Python LinearSVC implementation, which is part of the scikit-learn 0.19 library (Pedregosa et al., 2011), with the following parameter setting: class_weights=balanced, dual=False, max_iter=10000, and tol=1e-4.

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Using the SVM model formulated as above, we first tested how many genes are needed for accurate cell mapping. To this end, we randomly subset 20, 40, 60, and 80 genes from the list of differentially expressed genes and, for each gene set, built a vanilla SVM classification model to map each cell in the single-cell RNAseq dataset to its corresponding cell-type. The cross-validation accuracy was evaluated by using 4-fold cross-validation. Our results indicated that a high accuracy (>90%) can be obtained with 40 or more genes.

To map cell-types in the seqFISH data, we made a few modifications to incorporate the platform differences. First, since 125 genes were profiled by seqFISH, we used the top differentially expressed genes (P<1e-20) in the scRNAseq dataset for cell-type mapping. Based on the subsampling analysis described above, these 43 genes were sufficient for accurate cell-type mapping. Second, the scRNAseq data were z-score transformed so that the dynamic range was comparable with seqFISH data. Third, we used quantile normalization (Bolstad et al., 2003) to convert seqFISH data so that the statistical distribution was almost identical to single-cell RNAseq data. Fourth, we chose the regularization parameter *C* to maximize the cross-platform correlation between the cell-type specific gene expression profiles, resulting an estimate of *C*=1e-6. Finally, to account for the possibility that certain cells cannot be unequivocally assigned to a single cell-type, we used Platt scaling (ref) to convert SVM output to a probability measure and then selected a cutoff value of 0.5 probability to filter cells that can be confidently mapped to a single cell-type. 97 (5%) cells did not pass this filter.

Hidden Markov random field

Hidden Markov random field (HMRF) is a graph-based model commonly used for pattern recognition in image data analyses (Li, 2003; Zhang et al., 2001). In a common setting, HMRF is used to model the spatial distribution of a signal, such as the pixel intensities over a 2D image. The spatial structure is represented as a set of nodes on a regular grid, where neighboring nodes are connected to each other. The spatial pattern is "hidden" in the sense that it must be indirectly estimated from other variables that can be directly measured. The most important assumption in HMRF is the Markov property, which states that the spatial constraints can be reduced to considering only correlation between immediate neighboring nodes. This simplifying assumption implies that the joint distribution can be decomposed as products of much smaller components each defined on a fully connected subgraph (termed cliques). As has been done previously, we decomposed the graph into size-2 components (or edges in the graph) that provides a convenient means to estimating the MRF by using pairwise energies.

Specifically, let $S = \{s_i\}$ be the nodes in the graph. The set of nodes and the adjacency relation as defined by the local neighborhood graph forms the neighborhood system $(S, \{N_i\})$. Every node is associated with observed signal values x_i. Let $C = \{c_i = 1, \dots, K\}$ represent the set of

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

possible classes of patterns. The joint probability that a node s_i is associated with class c_i is specified by the following equation:

$$P(c_i|x_i, s_i, c_{N_i}) = 1/Z_1 \; P(x_i|c_i, s_i)P(c_i|s_i, c_{N_i}) \qquad \text{Eq. 2}$$

In the right-hand side, the term $P(x_i|c_i, s_i)$ models the effect of the node s_i's own gene expression, whereas $P(c_i|s_i, c_{N_i})$ models the effect of the neighboring cells configuration c_{N_i}. The combined effect of these two terms is schematically shown in **Fig. 2**. The latter term is further determined by the Gibbs distribution:

$$P(c_i|s_i, c_{N_i}) = 1/Z_2 \; exp\left(-\beta \sum_{s_j \in N_i} U(c_j, c_i)\right) \qquad \text{Eq. 3}$$

where $U(c_j, c_i)$ is referred to as the energy function. The exact formulation of $U(c_j, c_i)$ is dependent on the specific application, and it imposes the assumption of how neighboring nodes are interacting with each other. Here we use the special case Pott's model.

$$U(c_j, c_i) = -1, \; if \; c_j = \; c_i \; ; \; and \; 0 \; otherwise. \qquad \text{Eq. 4}$$

which means that the effects of neighboring cells are additive. Essentially, $P(c_i|s_i, c_{N_i})$ expresses the total energies as a summation of pairwise interaction energies with neighbors. The parameter beta reflects the strength of interactions.

Application to seqFISH data

The HMRF model described above is naturally applicable to analyze seqFISH data. Here each class of patterns corresponds to a spatial domain. The observed signals are gene expression levels measured by seqFISH data, whose distribution is modeled as a multivariate Gaussian random variable. The application of HMRF to seqFISH data analysis involves the following 4 components. 1) Neighboring graph representation. 2) Gene selection. 3) Domain number selection, and 4) Implementation and model inference. The details of each component are described below.

1) Neighborhood graph representation. An undirected graph was constructed to represent the spatial relationship between the cells. Each node represents a cell, and each edge connects a pair of neighboring cells. The neighborhood size was chosen such as on average each cell has five neighboring cells.
2) Gene selection. We selected a subset of genes whose expression patterns tend to be spatially coherent based on the following analysis. For each gene *g*, cells were divided into two mutually exclusive sets, corresponding to high expression (denoted as L1) and low expression (denoted as L0) respectively, at the 90th percentile expression level cutoff. The spatial coherence of the gene was quantified as the Silhouette coefficient (Rousseeuw,

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

1987) of the spatial distance associated with these two cell sets. Specifically, the Silhouette coefficient is calculated as:

$$\mathcal{S}_g = 1/|L_1| \sum_{s_i \in L_1} (m_i - n_i)/max(m_i, n_i) \qquad \text{Eq. 5}$$

where for a given cell s_i in Set L_1, m_i is defined as the average distance between s_i and any cell in L_0, and n_i is defined as the average distance between s_i and any other cell in L_1. Here, we used the rank-normalized, exponentially transformed distance to quantify the local physical distance between two cells. For a pair of cells s_i and s_j, this distance is defined as $r(s_i, s_j) = 1 - p^{rank_d(s_i,s_j)-1}$where $rank_d(s_i, s_j)$ is the mutual rank (Obayashi and Kinoshita, 2011) of s_i and s_j in the vectors of Euclidean distances $\{Euc(s_i,*)\}$ and $\{Euc(s_j,*)\}$. Hence, this exponentially weighted function (Moffat and Zobel, 2008) is designed to give more emphasis on closely located cells and penalizing far-away cells' distance to a large number. p is a rank-weighting constant (0<p<1.0) set at 0.95. The statistical significance of $\mathcal{S}_g$ was evaluated by random permutation, and the genes associated with significant values of $\mathcal{S}_g$ (p-value < 0.05) were selected as spatially coherent.

3) Domain number selection. We used k-means clustering results as initialization for the HMRF domains. The value of k was selected based on the gap-statistics (Tibshirani et al., 2001).
4) Implementation and model inference
   The model parameters were inferred by using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). We developed a new implementation based on the MRITC R package (Feng et al., 2012) and GraphColoring Java package (Brélaz, 1979). The implementation contains modifications to accommodate arbitrary neighborhood graph topology. The domain assignment for each cell was determined by using *maximum a posteriori* estimation, which can be viewed as the equilibrium state of the energy function.

Domain-specific gene signatures

For each spatial domain, we identified a subset of genes that were significantly (p < 0.05) up-regulated in the domain compared to cells in other regions, using the two-sample, one-sided t-test. A metagene expression was defined as the average expression level for this gene subset. We determined domain gene signatures for glutamatergic cells across the HMRF domains (see **Supplementary Table 6**) by summarizing the activities of genes that are simultaneously associated with a specific domain and cell-type.

Analysis of spatial structure in the single-cell RNAseq data

In order to systematically characterize the spatial structure within a single-cell RNAseq data, we summarized the gene signature associated with each spatial domain as a metagene. There are

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

nine metagenes in total, corresponding to domain I1a, I1b, O1, O2, O3, O4, IS, I2, and I3 respectively (defined in **Supplementary Table 6**). The overall activity of a metagene in each cell was quantified as the mean z-scored expression of all constituent genes in the signature and further binarized based on the bimodality of the distribution. A t-SNE analysis was performed on this matrix using the Rtsne package with parameters pca_scale=T, perplexity=35. Cell subpopulations with similar metagene expression patterns were identified by K-means clustering analysis (K=9).

For each subpopulation discovered from metagene clustering above, we found differentially expressed (DE) genes for the population (2-sample t-test, unequal variance, $P<0.05$). With the DE genes, we carried out Gene Ontology enrichment analysis (using hypergeometric test) for each of 9 subpopulations to construct a functional enrichment profile in **Fig. 4** (hypergeometric test $P<0.05$, top 500 DE genes analyzed per group, multiple hypothesis corrected by q-value procedure (Storey and Tibshirani, 2003)). Here we used genes expressed in glutamatergic cells as the background gene-set when doing enrichment analysis.

Tasic et al also provides layer information for a glutamatergic cell subset based on the layer from which the cells were manually dissected using different Cre-lines. To test whether the extracted subpopulation based on metagenes is enriched for a certain manually dissected layer of cells, we also performed hypergeometric test corrected for multiple hypothesis comparing manual annotations of cells to our HMRF domain-based annotations.

Code Availability

Code is deposited at https://bitbucket.org/qzhu/smfish-hmrf.

Data Availability

Expression data, spatial coordinates, SVM prediction results and HMRF segmentation results are deposited at https://bitbucket.org/qzhu/smfish-hmrf.

Ethical Compliance

Animal research was conducted in compliance with all relevant ethical regulations and other institutional requirements.

References

Achim, K., Pettit, J.B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509.

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Andjelic, S., Gallopin, T., Cauli, B., Hill, E.L., Roux, L., Badr, S., Hu, E., Tamas, G., and Lambolez, B. (2008). Glutamatergic Nonpyramidal Neurons From Neocortical Layer VI and Their Comparison With Pyramidal and Spiny Stellate Neurons. J. Neurophysiol. *101*, 641–654.

Bolstad, B.M., Irizarry, R. a, Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Ben Haim, L., and Rowitch, D.H. (2016). Functional diversity of astrocytes in neural circuit regulation. Nat. Rev. Neurosci. *18*, 31–41.

Brélaz, D. (1979). New methods to color the vertices of a graph. Commun. ACM *22*, 251–256.

Caicedo, J.C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A.S., Barry, J.D., Bansal, H.S., Kraus, O., et al. (2017). Data-analysis strategies for image-based cell profiling. Nat. Methods *14*, 849–863.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 348(6233): aaa6090.

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. Mach. Learn. *20*, 273–297.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P.S., Rothenberg, M.E., Leyrat, A.A., Sim, S., Okamoto, J., Johnston, D.M., Qian, D., et al. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol. *29*, 1120–1127.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Ofthe R. Stat. Soc. Ser. B *39*, 1–38.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. *343*, 193–196.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. J. Mach. Learn. Res. *9*, 1871–1874.

Feng, D., Tierney, L., and Magnotta, V. (2012). MRI Tissue Classification Using High-Resolution Bayesian Hidden Markov Normal Mixture Models. J. Am. Stat. Assoc. *107*, 102–119.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278.

Halpern, K.B., Shenhav, R., Matcovitch-Natan, O., Tóth, B., Lemze, D., Golan, M., Massasa, E.E., Baydatch, S., Landen, S., Moor, A.E., et al. (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature *542*, 1–5.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. (2011).

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011 Jul;21(7):1160-7.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. Science. *343*, 776–779.

Jaenisch R, Bird A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003 Mar;33 Suppl:245-54.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science. *358*, 194–199.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May 21;161(5):1187-1201.

Li, S.Z. (2003). Modeling image analysis problems using Markov random fields. in Stochastic Processes: Modelling and Simulation. 473-513.

Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods. 2014 Apr;11(4):360-1.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202-1214.

Moffat, A., and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. *27*, 1–27.

Moffitt, J.R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., and Zhuang, X. (2016). High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proc. Natl. Acad. Sci. *113*, 14456–14461.

Niu, W., Zang, T., Smith, D.K., Vue, T.Y., Zou, Y., Bachoo, R., Johnson, J.E., and Zhang, C.L. (2015). SOX2 reprograms resident astrocytes into neural progenitors in the adult brain. Stem Cell Reports *4*, 780–794.

Obayashi, T., and Kinoshita, K. (2011). COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res. *39*, D1016–D1022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2825–2830.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. *10*, 61–74.

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P.J., Carninci, P., Clatworthy, M., et al. (2017). Science Forum: The Human Cell Atlas. Elife *6*, e27041.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016a). In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron *92*, 342–357.

Shah, S., Lubeck, E., Schwarzkopf, M., He, T.-F., Greenbaum, A., Sohn, C.H., Lignell, A., Choi, H.M.T., Gradinaru, V., Pierce, N.A., et al. (2016b). Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. Development *143*, 2862–2867.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. *41*. D996-D1008.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. *99*, 12795–12800.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol.) *63*, 411–423.

Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. (2017). Challenges and emerging directions in single-cell analysis. Genome Biol. *18*.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. IEEE Transactions on Medical Imaging *20*, 45–57.

Zhang, Y.Y., Sloan, S.A.S.A.A., Clarke, L.E.L.E.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., et al. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron *89*, 37–53.

Figures

Figure 1.

Overall goal of the project and cell type prediction in seqFISH data.

a. Cellular heterogeneity is driven by both cell-type (indicated by shape) and environmental factors (indicated by colors). ScRNAseq based studies can only detect cell-type related variation, because spatial information is lost.
b. Our goal is to decompose the contributions of each factor by developing methods to integrate scRNAseq and seqFISH data.
c. Prediction results evaluated by the comparison of cell-type average expression profile across technologies for 8 major cell types. Values represent expression z-scores. Genes are ordered by significance of differential expression in scRNAseq.
d. Correlation between reference and predicted cell type averages ranges from 0.75 to 0.95.
e. Integration of seqFISH and scRNAseq data (illustrated by b) enables cell-type mapping with spatial information in the adult mouse visual cortex. Each cell type is labeled by a different color. Cell shape information is obtained from segmentation of cells from images (see Methods).

Figure 2.

Spatial domain dissection in seqFISH data using hidden Markov random field (HMRF) approach.

a. A schematic overview of the HMRF model. A neighborhood graph represents the spatial relationship between imaged cells (indicated by the circles) in the seqFISH data. The edges connect cells that are neighboring to each other. seqFISH-detected multigene expression profiles are used together with the graph topology to identify spatial domains. In contrast, k-means and other clustering methods do not utilize spatial information therefore the results are expected to be less coherent (illustrated in the dashed box).

bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

b. An intuitive illustration of the basic principles in a HMRF model. For a hypothetical cell (indicated by the question mark), its spatial domain assignment is inferred from combining information from gene expression (x_i) and neighborhood configuration (c_{N_i}). The color of each node represents cell's expression and the number inside each node is domain number. In this hypothetical example, combining such information results the cell being assigned to domain 1, instead of domain 3 (see Methods).
c. HMRF identifies spatial domain configuration in the mouse visual cortex region. Distinct domains reveal a resemblance to layer organization or cortex. Naming of domains: I1a, I1b, I2, I3 are inner domains distributed in the inner layers. O1-O4 are outer domains. IS is inner scattered state. These domains are associated with cell morphological features such as distinct cell shape differences in outer layer domains. Cell shape information is obtained from segmentation of cells from images (see Methods).
d. General domain signatures that are shared between cells within domains.

Figure 3.

HMRF analysis identified domain associated heterogeneity within glutamatergic cells.

a. Three major sources of variations in glutamatergic neurons. (Top): cell type specific signals Gda and Tbr1. (Middle): general domain signatures as in Fig 2d, summarized into metagenes' expression. (Bottom): glutamatergic specific domain signatures, found by comparing glutamatergic cells across domains and removing signatures that also vary across domains in other cell types.
b. Snapshots of single cells. Each row is a snapshot of cells at the boundary of two layers. Each of two columns is a type of annotation: (left column) cell type, (right column) HMRF domains. Cell type is incapable of explaining layer-to-layer morphological variations: e.g. glutamatergic cells (orange) is present in all layers yet morphological differences exist within glutamatergic cells. HMRF domains better capture the boundary of two layers in each case, in that the domains can separate distinct morphologies

Figure 4.

Reanalysis of single-cell RNAseq data (from Tasic *et al*) with domain signatures summarized into metagenes.

a. t-SNE plot shows how glutamatergic cells from Tasic et al cluster according to glutamatergic-specific domain signatures aggregated as metagenes (shown in (b)). Colors indicate k-means clusters (k=9). Each cluster is annotated by its enriched metagene activity.
b. Binarized metagene expression profiles for the glutamatergic cells. Red: population that highly expresses the metagene.

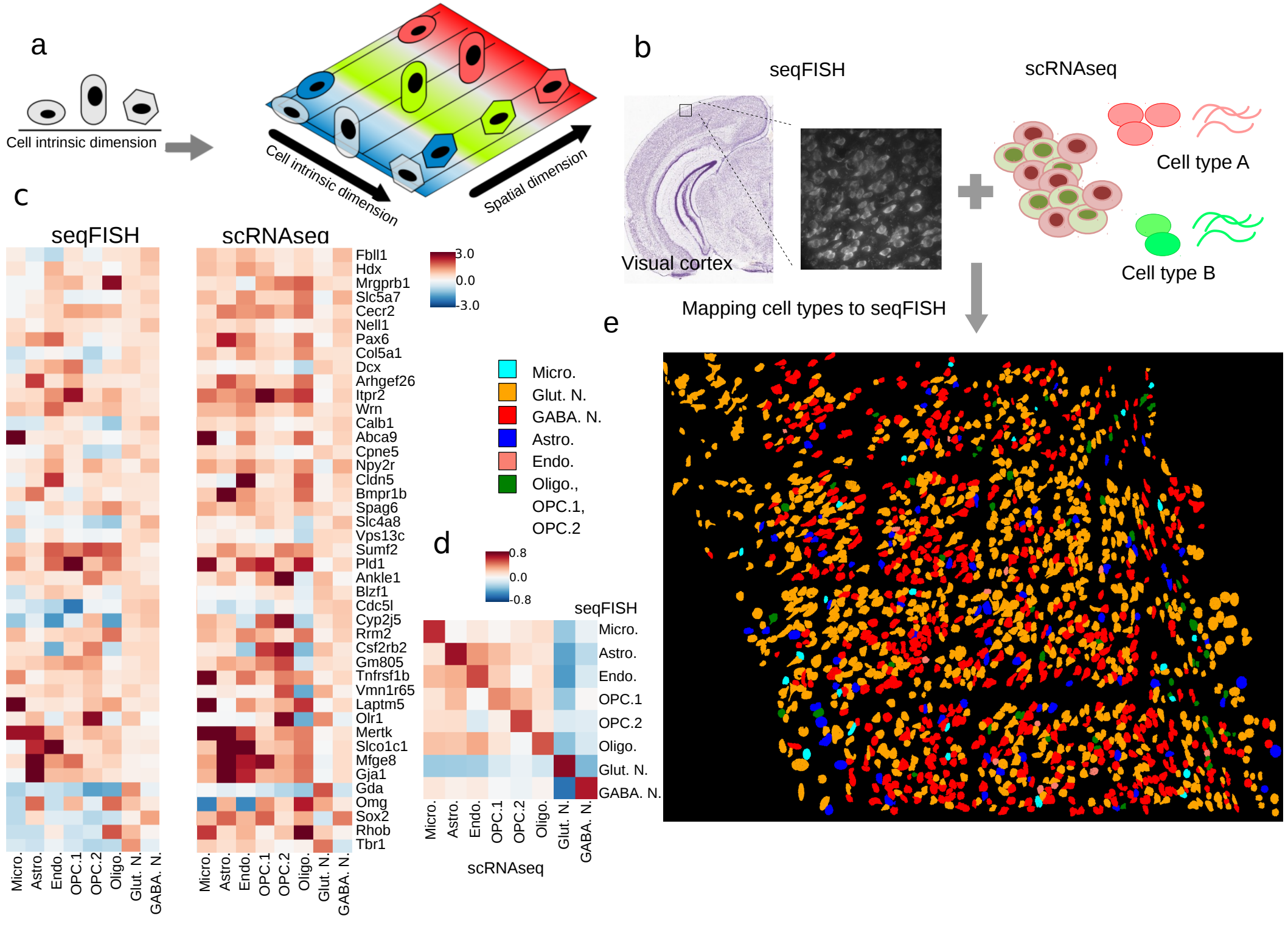
bioRxiv preprint first posted online Mar. 2, 2018; doi: http://dx.doi.org/10.1101/275156. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

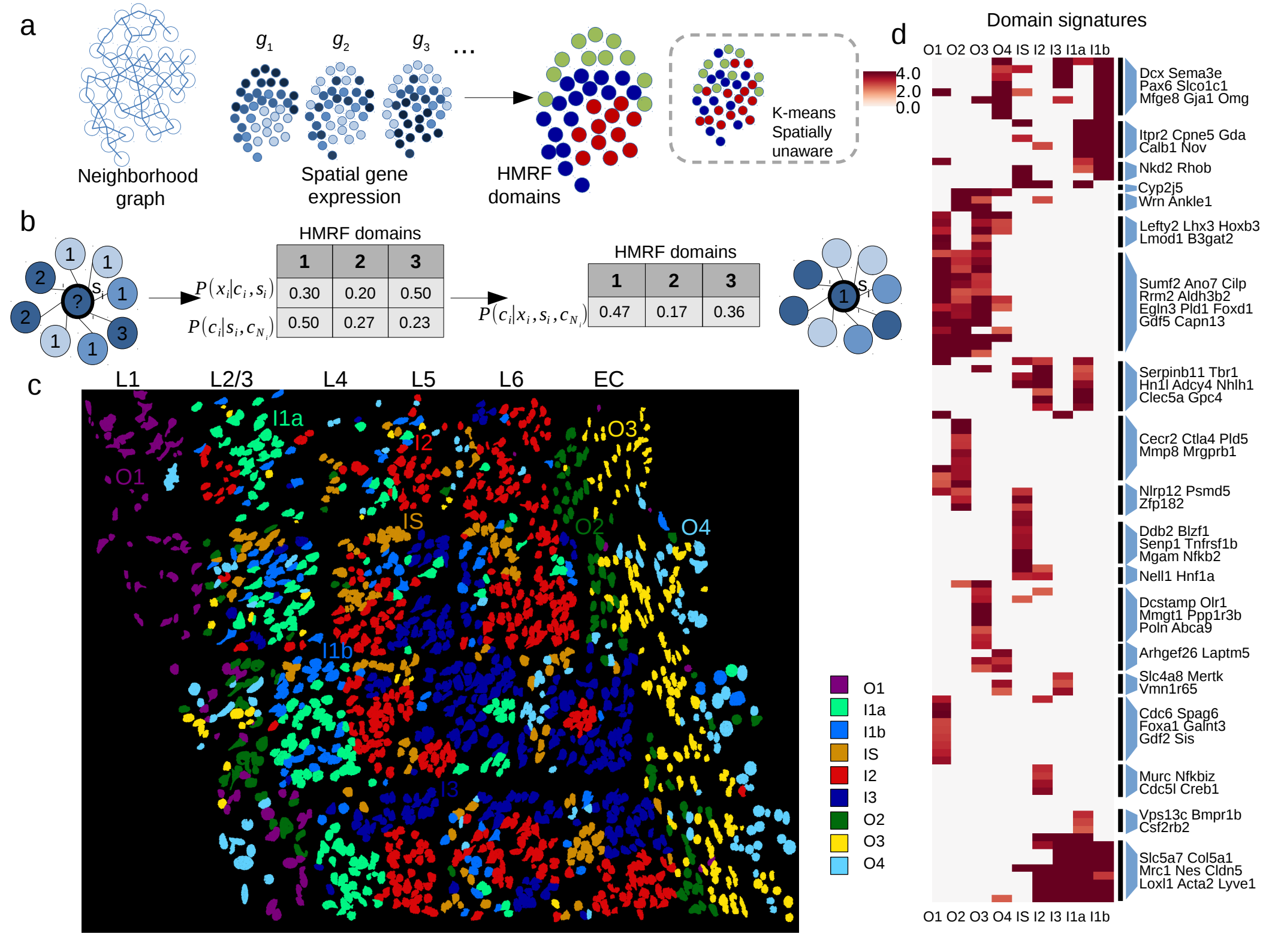
c. Spatial clusters defined according metagenes are enriched in manual layer dissection annotations. Column: layer information obtained from microdissection. Row: metagene based cell clusters.
d. Inferred spatial clusters of glutamatergic neurons are enriched in distinct GO biological processes.

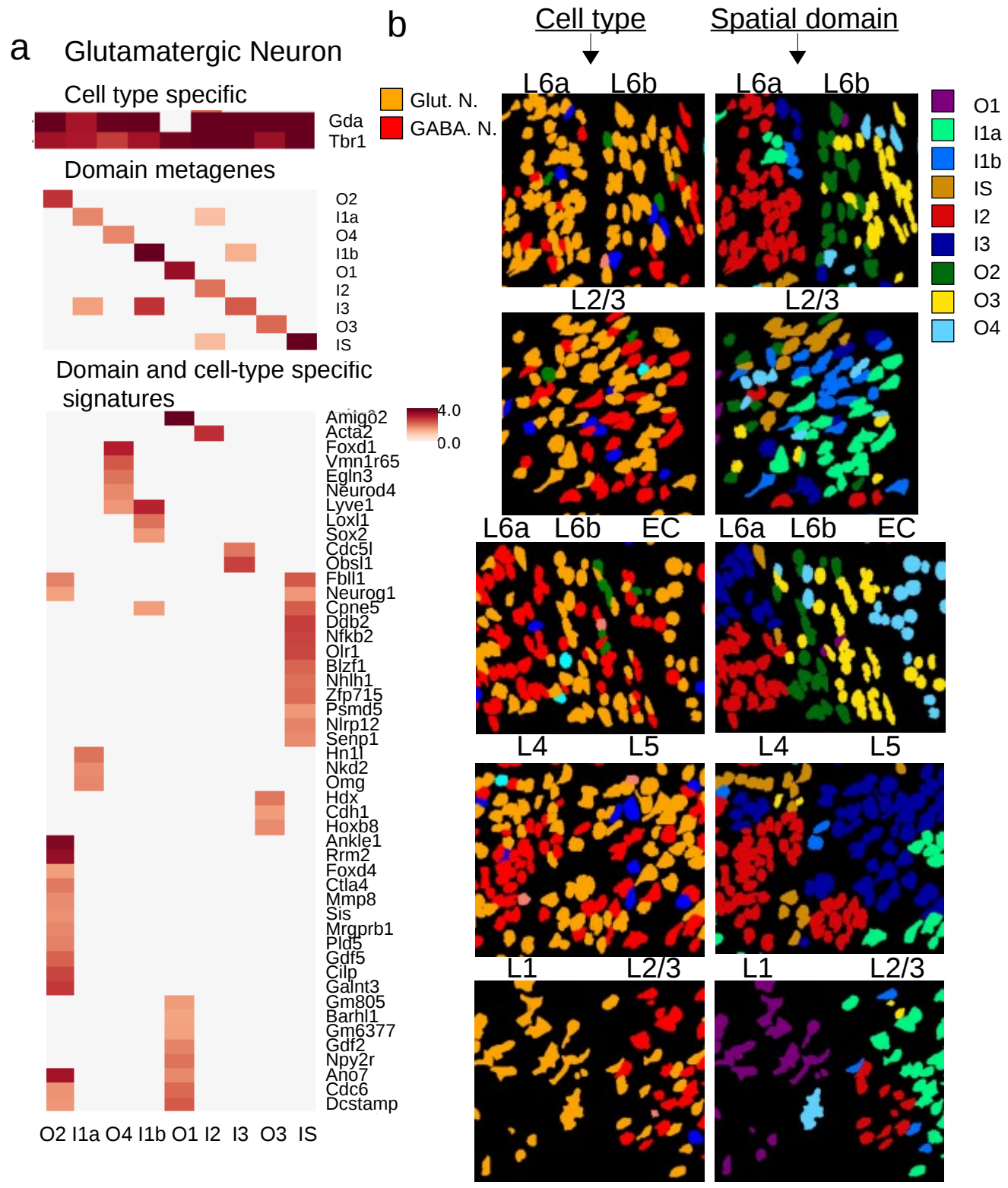
Figure 5.

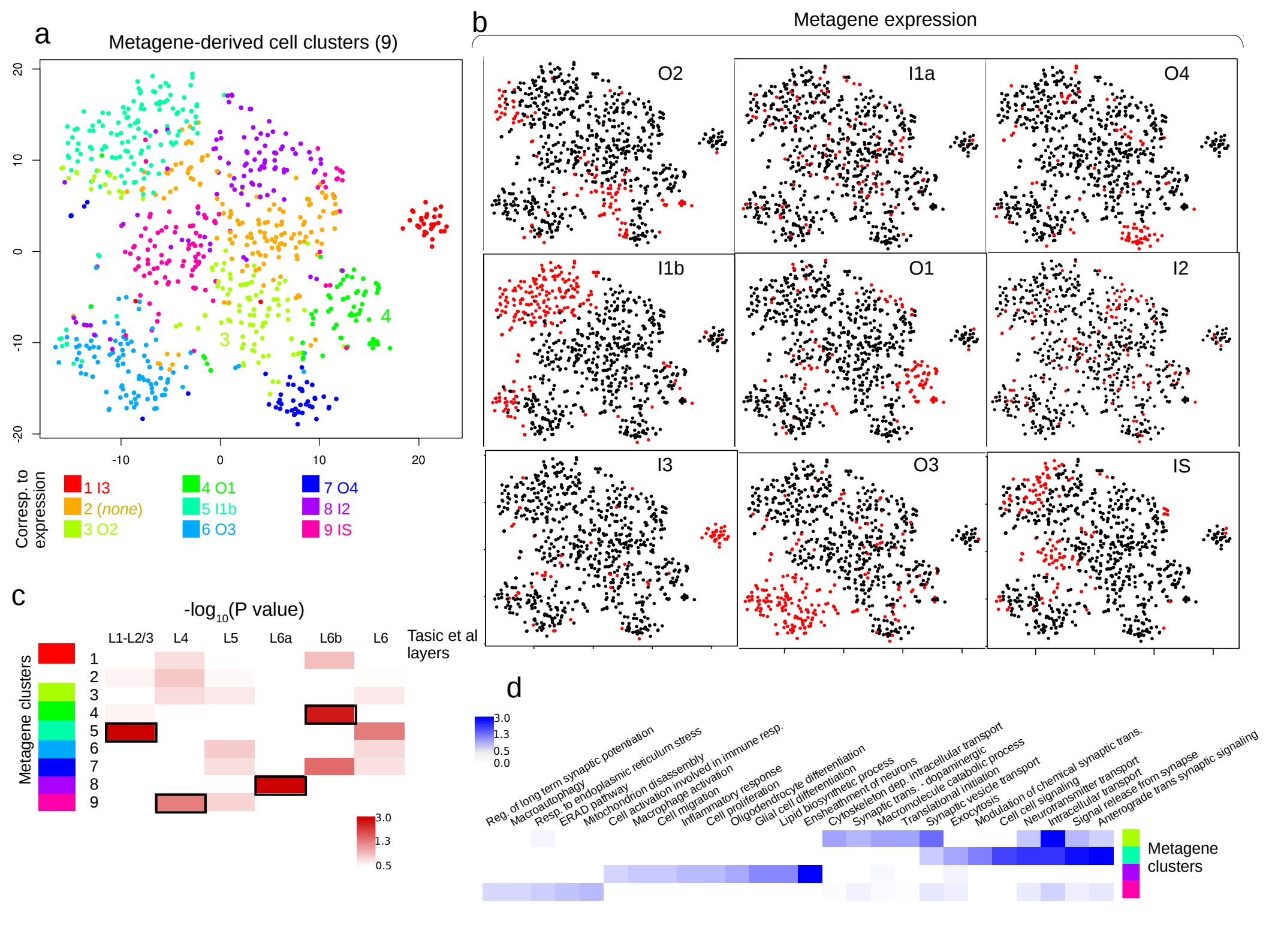
Spatially dependent astrocyte variation revealed by HMRF.

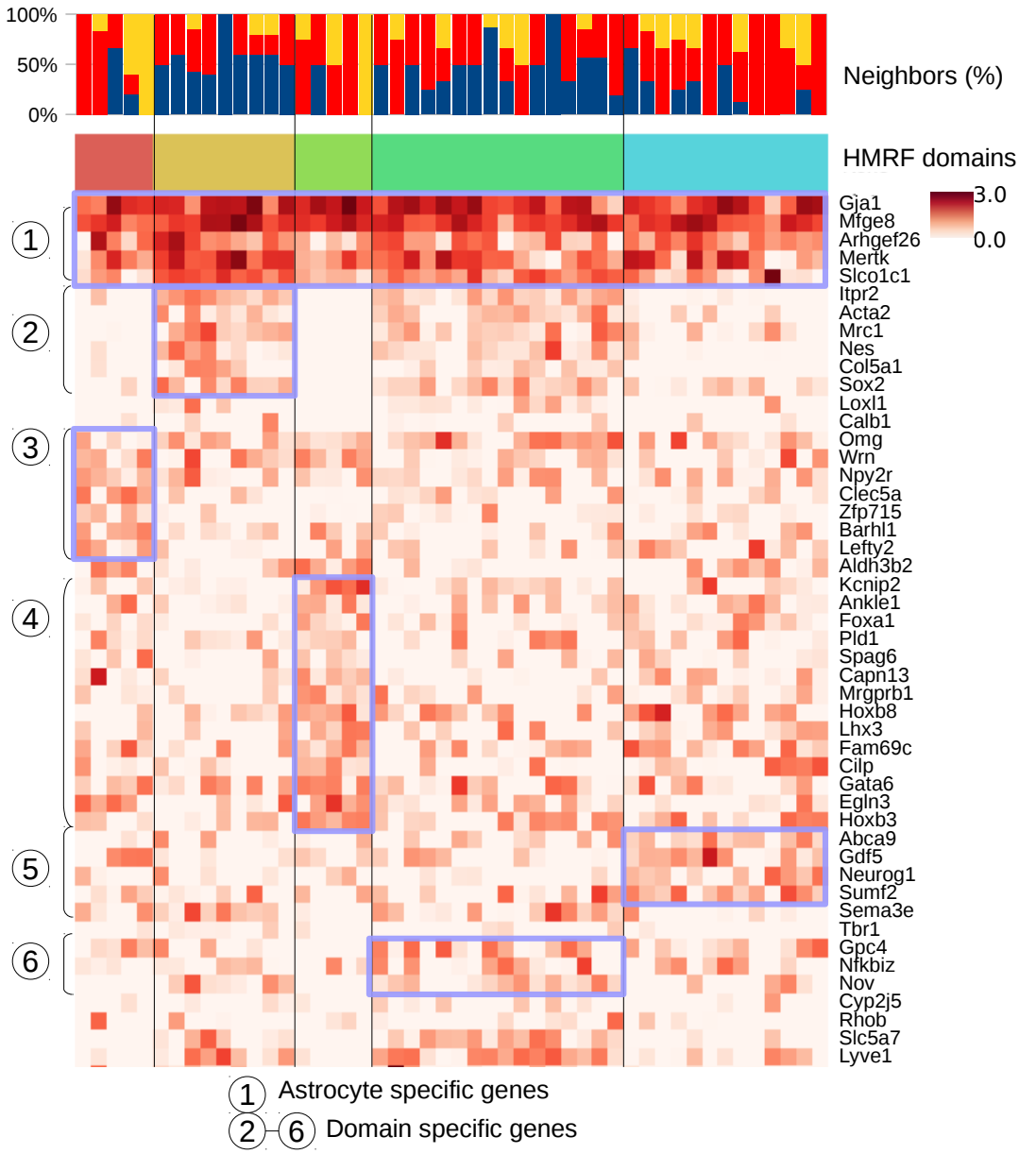
Neighborhood cell type composition for the 47 astrocyte cells (columns). Cells are ordered by HMRF domain annotations. The heatmap shows single cell expression of astrocytes clustered by domain-specific genes. Blue-box highlights the common signatures expressed in each domain's astrocyte population.


a
Cell intrinsic dimension
Cell intrinsic dimension
Spatial dimension
b
seqFISH
scRNAseq
Visual cortex
Cell type A
Cell type B
Mapping cell types to seqFISH
c
seqFISH
scRNAseq
3.0
0.0
-3.0
Fbll1
Hdx
Mrgprb1
Slc5a7
Cecr2
Nell1
Pax6
Col5a1
Dcx
Arhgef26
Itpr2
Wrn
Calb1
Abca9
Cpne5
Npy2r
Cldn5
Bmpr1b
Spag6
Slc4a8
Vps13c
Sumf2
Pld1
Ankle1
Blzf1
Cdc5l
Cyp2j5
Rrm2
Csf2rb2
Gm805
Tnfrsf1b
Vmn1r65
Laptm5
Olr1
Mertk
Slco1c1
Mfge8
Gja1
Gda
Omg
Sox2
Rhob
Tbr1
Micro.
Astro.
Endo.
OPC.1
OPC.2
Oligo.
Glut. N.
GABA. N.
d
0.8
0.0
-0.8
seqFISH
scRNAseq
e
Micro.
Glut. N.
GABA. N.
Astro.
Endo.
Oligo.,
OPC.1,
OPC.2


a
Neighborhood graph
g_1 g_2 g_3 ...
Spatial gene expression
HMRF domains
K-means
Spatially unaware
b
HMRF domains
1 2 3
$P(x_i|c_i,s_i)$ 0.30 0.20 0.50
$P(c_i|s_i,c_{N_i})$ 0.50 0.27 0.23
HMRF domains
1 2 3
$P(c_i|x_i,s_i,c_{N_i})$ 0.47 0.17 0.36
c
L1 L2/3 L4 L5 L6 EC
O1 I1a I2 O3 IS O2 O4 I1b I3
O1 I1a I1b IS I2 I3 O2 O3 O4
d
Domain signatures
O1 O2 O3 O4 IS I2 I3 I1a I1b
4.0 2.0 0.0
Dcx Sema3e Pax6 Slco1c1 Mfge8 Gja1 Omg
Itpr2 Cpne5 Gda Calb1 Nov
Nkd2 Rhob
Cyp2j5
Wrn Ankle1
Lefty2 Lhx3 Hoxb3 Lmod1 B3gat2
Sumf2 Ano7 Cilp Rrm2 Aldh3b2 Egln3 Pld1 Foxd1 Gdf5 Capn13
Serpinb11 Tbr1 Hn1l Adcy4 Nhlh1 Clec5a Gpc4
Cecr2 Ctla4 Pld5 Mmp8 Mrgprb1
Nlrp12 Psmd5 Zfp182
Ddb2 Blzf1 Senp1 Tnfrsf1b Mgam Nfkb2
Nell1 Hnf1a
Dcstamp Olr1 Mmgt1 Ppp1r3b Poln Abca9
Arhgef26 Laptm5
Slc4a8 Mertk Vmn1r65
Cdc6 Spag6 Foxa1 Galnt3 Gdf2 Sis
Murc Nfkbiz Cdc5l Creb1
Vps13c Bmpr1b Csf2rb2
Slc5a7 Col5a1 Mrc1 Nes Cldn5 Loxl1 Acta2 Lyve1
O1 O2 O3 O4 IS I2 I3 I1a I1b


a
Glutamatergic Neuron
Cell type specific
Gda
Tbr1
Glut. N.
GABA. N.
Domain metagenes
O2
I1a
O4
I1b
O1
I2
I3
O3
IS
Domain and cell-type specific signatures
4.0
0.0
Amigo2
Acta2
Foxd1
Vmn1r65
Egln3
Neurod4
Lyve1
Loxl1
Sox2
Cdc5l
Obsl1
Fbll1
Neurog1
Cpne5
Ddb2
Nfkb2
Olr1
Blzf1
Nhlh1
Zfp715
Psmd5
Nlrp12
Senp1
Hn1l
Nkd2
Omg
Hdx
Cdh1
Hoxb8
Ankle1
Rrm2
Foxd4
Ctla4
Mmp8
Sis
Mrgprb1
Pld5
Gdf5
Cilp
Galnt3
Gm805
Barhl1
Gm6377
Gdf2
Npy2r
Ano7
Cdc6
Dcstamp
O2 I1a O4 I1b O1 I2 I3 O3 IS
b
Cell type
Spatial domain
L6a L6b
L6a L6b
L2/3
L2/3
L6a L6b EC
L6a L6b EC
L4 L5
L4 L5
L1 L2/3
L1 L2/3
O1
I1a
I1b
IS
I2
I3
O2
O3
O4


a
Metagene-derived cell clusters (9)
20
10
0
-10
-20
-10
0
10
20
4
3
Corresp. to expression
1 I3
2 (none)
3 O2
4 O1
5 I1b
6 O3
7 O4
8 I2
9 IS
b
Metagene expression
O2
I1a
O4
I1b
O1
I2
I3
O3
IS
c
-log10(P value)
L1-L2/3
L4
L5
L6a
L6b
L6
Tasic et al layers
Metagene clusters
1
2
3
4
5
6
7
8
9
3.0
1.3
0.5
d
3.0
1.3
0.5
0.0
Reg. of long term synaptic potentiation
Macroautophagy
Resp. to endoplasmic reticulum stress
ERAD pathway
Mitochondrion disassembly
Cell activation involved in immune resp.
Macrophage activation
Cell migration
Inflammatory response
Cell proliferation
Oligodendrocyte differentiation
Glial cell differentiation
Lipid biosynthetic process
Ensheathment of neurons
Cytoskeleton dep. intracellular transport
Synaptic trans. - dopaminergic
Macromolecule catabolic process
Translational initiation
Synaptic vesicle transport
Exocytosis
Modulation of chemical synaptic trans.
Cell cell signaling
Neurotransmitter transport
Intracellular transport
Signal release from synapse
Anterograde trans synaptic signaling
Metagene clusters


100%
50%
0%
Neighbors (%)
HMRF domains
3.0
0.0
Gja1
Mfge8
Arhgef26
Mertk
Slco1c1
Itpr2
Acta2
Mrc1
Nes
Col5a1
Sox2
Loxl1
Calb1
Omg
Wrn
Npy2r
Clec5a
Zfp715
Barhl1
Lefty2
Aldh3b2
Kcnip2
Ankle1
Foxa1
Pld1
Spag6
Capn13
Mrgprb1
Hoxb8
Lhx3
Fam69c
Cilp
Gata6
Egln3
Hoxb3
Abca9
Gdf5
Neurog1
Sumf2
Sema3e
Tbr1
Gpc4
Nfkbiz
Nov
Cyp2j5
Rhob
Slc5a7
Lyve1
1
2
3
4
5
6
1 Astrocyte specific genes
2–6 Domain specific genes